Weather and Climate Extremes xxx (2017) 1–12

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

Quantifying statistical uncertainty in the attribution of human influence on severe weather

Christopher J. Paciorek [a,*], Dáithí A. Stone [b], Michael F. Wehner [b]

[a] Department of Statistics, University of California, Berkeley CA 97420, USA
[b] Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA

ARTICLE INFO

Keywords:
Event attribution
Climate change
Uncertainty quantification
Likelihood ratio
Bootstrap
Confidence interval

ABSTRACT

Event attribution in the context of climate change seeks to understand the role of anthropogenic greenhouse gas emissions on extreme weather events, either specific events or classes of events. A common approach to event attribution uses climate model output under factual (real-world) and counterfactual (world that might have been without anthropogenic greenhouse gas emissions) scenarios to estimate the probabilities of the event of interest under the two scenarios. Event attribution is then quantified by the ratio of the two probabilities. While this approach has been applied many times in the last 15 years, the statistical techniques used to estimate the risk ratio based on climate model ensembles have not drawn on the full set of methods available in the statistical literature and have in some cases used and interpreted the bootstrap method in non-standard ways. We present a precise frequentist statistical framework for quantifying the effect of sampling uncertainty on estimation of the risk ratio, propose the use of statistical methods that are new to event attribution, and evaluate a variety of methods using statistical simulations. We conclude that existing statistical methods not yet in use for event attribution have several advantages over the widely-used bootstrap, including better statistical performance in repeated samples and robustness to small estimated probabilities. Software for using the methods is available through the climextRemes package available for R or Python. While we focus on frequentist statistical methods, Bayesian methods are likely to be particularly useful when considering sources of uncertainty beyond sampling uncertainty.

1. Introduction

Over the past decade, there has been increasing interest in the climate change research community in describing the role of anthropogenic greenhouse gas emissions in specific extreme weather events, commonly referred to as "event attribution" (Stott et al., 2013; National Academies of Sciences, Engineering, and Medicine, 2016; Herring et al., 2016). This increased scientific interest has been motivated by public interest, as the global warming signal both becomes more noticeable and easier to analyse, and an expectation that further understanding of the role of anthropogenic emissions might inform adaptation activities. Concurrently, observationally-based data products, numerical climate models, and computational resources have developed to the point where they can be usefully applied to the analysis of long-term trends in extreme weather.

Allen (2003) first noted the potential public demand for event attribution information and suggested that concepts from epidemiology and environmental law would be appropriate in the climate change setting as well. In this setting, one compares the probability of an extreme weather event under a factual scenario of recent and current conditions to the probability in a counterfactual scenario in which anthropogenic emissions had never occurred but other factors (e.g., the eruption of Mt. Pinatubo) had still influenced the climate system. Stone and Allen (2005) formally introduced the concepts of fraction attributable risk (FAR) and risk ratio (RR) and proposed how they might be estimated given available tools. (Note that 'risk' in risk ratio inherits from its usage in biostatistics/epidemiology and is unrelated to statistical risk.) Unlike in epidemiology, in which repeated observations (e.g., multiple patients) are available with different exposures to potential health risks, in the climate context we do not have available repeated samples of the world, particularly under the counterfactual scenario. Analyses therefore use simulations of numerical models of the climate system as surrogates. Stott et al. (2004) presented the first study using this approach, making use of a small number of simulations of a climate model representing the entire climate system, but the small number of simulations meant that they had to assume a relationship between the average summer

* Corresponding author.
E-mail address: paciorek@stat.berkeley.edu (C.J. Paciorek).

https://doi.org/10.1016/j.wace.2018.01.002
Received 15 June 2017; Received in revised form 28 December 2017; Accepted 10 January 2018
Available online xxxx

temperature and the frequency of extremely hot summers. Pall et al. (2011) extended the method to take advantage of an exceptionally large number of simulations that removed the need for the assumption of a mean-extreme relationship, at the expense of an incomplete model of the climate system. Together, these two variants of the "probabilistic event attribution" method have become the most popular approach for event attribution research in recent years (e.g., Peterson et al., 2012, 2013; Herring et al., 2014, 2015, 2016).

Despite this popularity, there has been little formal statistical development of the technique since Stone and Allen (2005). The exception is the work of Hansen et al. (2014), who considered the problem of estimating the FAR as the ratio of means of two independent Poisson processes and proposed use of a confidence interval well-established in the statistical literature (Wilson, 1927). They also considered the case when the estimated counterfactual probability of an event is zero and developed a one-sided interval for FAR.

Underlying probabilistic event attribution are the existence of observational data and relevant climate model simulations, along with the assumption that both are adequate for performing an analysis. Guidelines for such adequacy tests are still being considered (Stott et al., 2004; Pall et al., 2011; Lott and Stott, 2016; Angélil et al., 2016). We leave the question of how to evaluate the input data sources for further discussion elsewhere and focus instead on statistical methods to estimate the RR (conditional on the data sources) and to quantify the uncertainty in the estimate due to limited statistical sampling because of finite model ensemble sizes.

This paper presents a formal frequentist statistical framework for probabilistic event attribution. Previous work has frequently involved interpretation of frequentist-based analysis using a Bayesian perspective (e.g., Stott et al., 2004; Pall et al., 2011), so part of our goal is to clarify ways to implement standard frequentist statistical methods in this context. We focus on estimating uncertainty from limited sampling, for which a frequentist approach is well-suited. We present several statistical methods for estimating uncertainty via confidence intervals, assess performance in a statistical simulation study, provide an example analysis, and make recommendations for the practice of probabilistic event attribution. We highlight that some methods allow estimation of a confidence interval even when the estimated counterfactual probability is zero and the risk ratio is infinity.

2. Statistical framework

We consider the approach of estimating the risk (probability) ratio, as the ratio of the probability of a specified event under a factual scenario (F), p_F, to that probability under a counterfactual scenario (CF), p_C:

$$\mathrm{RR} = \frac{p_F}{p_C}.$$

In most climate attribution studies (and a case study in Section 6), the factual scenario is intended to represent the climate we have experienced, i.e., in which historical emissions of greenhouse gases and other drivers of climate change occurred as they did. In contrast, the counterfactual is intended to represent a climate that might have been in the absence of human interference, i.e., in which the anthropogenic drivers are held at some "pre-industrial" level (e.g., year 1850 values) even as the natural drivers (volcanic eruptions and solar luminosity) remain varying as we have experienced. Given estimates, $\widehat{p}_F$ for p_F and $\widehat{p}_C$ for p_C, one simply estimates $\widehat{\mathrm{RR}} = \frac{\widehat{p}_F}{\widehat{p}_C}$. We focus on RR rather than FAR because of its use in epidemiology and related fields, its interpretability, and because in our statistical development we generally work with log (RR), which is expressed simply as the difference in log probabilities and which improves statistical performance. Usefully, on the log scale, increases and decreases from anthropogenic influence are symmetric about zero, unlike for the FAR. However, a confidence interval for FAR can be trivially calculated from a confidence interval for RR.

In this work we lay out a classical (or frequentist) statistical framework that treats p_F and p_C (and therefore RR) as fixed, non-random quantities that are properties of the climate system. Uncertainty arises in estimating these quantities, and our goal is to quantify the uncertainty in $\widehat{\mathrm{RR}}$ as an estimate of the unknown RR based on the (sampling) probability density of $\widehat{\mathrm{RR}}$. In contrast, a Bayesian treatment would consider these quantities to be random variables with probability densities and would allow one to make probabilistic statements directly about RR. Note that in previous event attribution work, researchers have frequently interpreted results, including bootstrap-based analyses, in a Bayesian fashion despite largely relying on frequentist methods (e.g., Stott et al., 2004; Pall et al., 2011).

2.1. Basic probabilistic framework

We use R to denote the continuous variable of interest (e.g., runoff or rainfall) and $I(R > c)$ to denote the occurrence of the event of interest, defined by whether the variable exceeds some cutoff, c. $I(\cdot)$ is the indicator function that is 1 if the condition occurs and 0 if not.

Let

$$p = P(R > c) = E(I(R > c)) = \int I(R > c) f(r) dr \tag{1}$$

be the probability of the event, where $E(\cdot)$ denotes expected value and $f(r)$ is the probability density function of R. While in principle we can estimate this quantity in the real world using observations, we only have a single observational series. If that series were stationary (which may be a reasonable assumption only for short time periods), we might use multiple years as replicates, but we cannot estimate this quantity under the counterfactual scenario. Therefore researchers rely on climate model simulations to estimate $p \in \{p_F, p_C\}$. (However the methods in this paper can be used for a RR contrasting probabilities under two different time periods.)

In a single climate model simulation, R for a pre-specified time and location is not random since the model is deterministic and therefore R (and $I(R > c)$) is a fixed value. In this case, the use of expectation and probability above is not meaningful. However, not only can models be run under multiple scenarios, but they can be run multiple times with each realization differing in the initial state and thus the subsequent weather produced. Let W (for "weather") be a (very high-dimensional) random variable indicating the state of the earth system and $f(w)$ the probability density function of W. The individual realizations provide a simple random sample, $w_1, \ldots, w_{n_w}$, of size n_w from $f(w)$. By using the initial condition ensemble, passed through the deterministic GCM, to represent $f(w)$, we rely on the strong assumption that use of the sample of initial conditions to initialize the GCM can approximate the true $f(w)$ under a given scenario. This ergodic assumption is reasonable for the atmospheric model simulations used in the approach of Pall et al. (2011) after a spin-up period of weeks to over a year but requires decades for atmosphere-ocean GCMs. W is random and induces the randomness in R. Thus $R(W)$ is our random variable of interest. Now we have

$$p = P(R(W) > c) = E(I(R(W) > c)) = \int I(R(w) > c) f(w) dw. \tag{2}$$

The discussion above assumes one value of R per model realization, which would often be the case for annual or seasonal extremes or for short-term extremes for specific calendar dates. If one has multiple values of R for a given realization (e.g., analyzing daily rainfall), one would generally benefit from having a much larger sample size than is the focus here (particularly the simulation results of Section 5), but one would need to account for any correlation between the multiple values of R from a given realization (e.g., the daily values in a given season or year), particularly when estimating uncertainty.

In Section 3 we present methods to estimate $p \in \{p_F, p_C\}$, while in Section 4 we discuss how to quantify the uncertainty in estimating RR. Before doing so, we consider the additional complexities involved in

estimating RR that are not characterized in the basic framework above.

2.2. Sources of uncertainty

Sources of uncertainty in estimating the RR using GCMs (see also National Academies of Sciences, Engineering, and Medicine (2016)) include:

- **Variability in the earth system:** Uncertainty arises from limited sampling of the variability of the system, which can range from time scales of days to years, decades, and centuries. This is often operationalized via an ensemble of GCM simulations initialized with different initial conditions and quantified based on the notion of sampling uncertainty as discussed in the basic framework above.
- **Boundary condition uncertainty**: This includes aspects of the system that can vary in time but that are prescribed in the model and thus not simulated by it, such as the concentrations of radiatively-active trace constituents in the atmosphere and changes in land use. Modification of one or more of these boundary conditions constitutes the distinction between the factual and counterfactual scenarios; thus uncertainty in the distinction is implicit in this source of uncertainty. When such forcing terms are based on observational data, there is observational uncertainty.
- **Model parametric uncertainty:** This represents uncertainty in the appropriate values for parameters in the GCM that are used in approximations for various processes not directly simulated by the GCM. This could include fundamental physical or chemical constants, but generally uncertainty in those constants is negligible relative to uncertainty in the appropriate value of bespoke parameters.
- **Model structural uncertainty:** This is another component of uncertainty inherent to the climate model – the uncertainty in how to represent a complex physical system as a mathematical model, but not the uncertainty in tuning of that approximation. Note that determination of whether a model is fit for purpose for event attribution is a binary determination of this broader component of uncertainty.

We draw a fundamental distinction between uncertainty arising from variability in the system and the remaining sources of uncertainty. Uncertainty from variability is a statistical sampling problem, with uncertainty decreasing with increasing sample size. Critically, this uncertainty is quantifiable using well-established frequentist or Bayesian statistical methods. If we had an infinitely large ensemble, our sampling uncertainty would be zero. However, there would still be bias from the fact that the climate produced by the model is not the same as the climate produced by the real system, even in equilibrium, due to the additional factors listed above. In this work, we focus on uncertainty from variability, in part because frequentist methods would appear to be of limited use for the other sources. We make some comments on these other components of uncertainty in the discussion.

Finally, many event attribution analyses use atmospheric-land GCMs, as in Pall et al. (2011), rather than coupled models, for reasons of computational efficiency, a partial conversion of (oceanic) model structural uncertainty to better-understood boundary condition uncertainty, and in some cases a desire to more strongly condition the analysis on known features of the real world. With such analyses, model simulations do not sample from the longer-time-scale internal variability of the system. In Supplemental Material B we describe the additional uncertainty related to this longer-time-scale variability and present statistical methods for averaging over results from multiple years.

3. Estimating event probabilities

We next provide an overview and comparison of methods for estimating $p \in \{p_F, p_C\}$. We discuss approaches that simply count exceedances and rely on binomial sampling statistics, parametric fitting of the variable of interest, and extreme value analysis (EVA) of the variable of interest.

While EVA is often used for estimating probabilities of extreme events with observational data, serious difficulties can arise in the context of event attribution analyses that use model ensembles, and simple nonparametric estimators are often a good choice. First note that EVA is generally used on long time series and applied to short-term (e.g., daily) extremes. In this context, one often uses a block maximum (or minimum) approach, blocking by year, or a peaks-over-threshold approach that only uses observations over a high threshold, such as the 99th percentile of the observations. The statistical theory that supports EVA relies on taking maxima (or minima) over a large number of observations per block or setting a high threshold. However, for analysis of annual or seasonal extremes or short-term extremes for specific calendar dates, model ensembles generally only provide moderate sample sizes such as 50, 100, or 400, violating the assumptions of EVA. Furthermore, in event attribution, the event of interest may be extreme only in one scenario, so EVA may not be applicable for one scenario. And for short model simulations with initial conditions that favor extreme weather, the entire sample may be extreme in an absolute sense, in which case the event of interest may, in a relative (i.e., conditional) sense, not be extreme and EVA would not be appropriate (e.g., Pall et al., 2017).

Thus, while there are situations in which EVA has advantages for model-based event attribution (as seen in the case study in Section 6 and discussed further in our recommendations in Section 7), we focus more on the binomial approach because of its greater generality.

3.1. Nonparametric (binomial sampling of events)

One straightforward approach to estimating p is a nonparametric Monte Carlo (MC) estimate based on the available sample. An MC estimate of the expectation in (2) involves drawing n_w samples of W from $f(w)$ (based on the model simulations) and using the estimator:

$$\widehat{p} = \frac{\#\text{events}}{n_w} = \frac{1}{n_w}\sum_{i=1}^{n_w} I(R(w_i) > c), \tag{3}$$

which is justified by the law of large numbers because $\widehat{p}$ is a statistically-consistent estimator for p as $n_w \to \infty$. Consistency here means that as n_w gets large, $\widehat{p}$ is guaranteed to get very close to p.

This estimator makes no assumptions about the distribution of R (hence the term 'nonparametric', although it is equivalent to Bernoulli sampling and a resulting binomial distribution for the number of events) and is thus robust, but there are drawbacks to the approach. First, the estimator may have more uncertainty than parametric estimators that assume a particular distribution. Furthermore, if $R > c$ does not occur in the sample, our estimator is $\widehat{p} = 0$ even when we have substantive expertise suggesting that p is non-zero. Zeros for $\widehat{p}_F$, $\widehat{p}_C$, or both result in RR estimates of zero, infinity, or an undefined value.

3.2. Parametric

If one is willing to assume a particular parametric form for the distribution of R, such as a normal or log-normal distribution, statistical theory tells us that one may be able to obtain an estimator for p that has lower variance (less uncertainty) than the nonparametric estimator above. For example if we assume normality, we can estimate the mean and variance of R as $\bar{r}$ and $s_R^2 = \frac{\sum (r_i - \bar{r})^2}{n_w - 1}$, where $r_i = R(w_i)$ and $\bar{r}$ is the simple average of $r_1, \ldots, r_{n_w}$. Then $\widehat{p} = \int_c^\infty f(r; \bar{r}, s_R^2) dr$ where $f(r; \bar{r}, s_R^2)$ is the normal density with mean $\bar{r}$ and variance s_R^2.

The methodology allows us to estimate probabilities for events far in the tail of the distribution, but it relies crucially on the assumption that the chosen distribution well approximates the true distribution even far in the tail. Put another way, all the data values are used to estimate the parameters of the assumed distribution and thereby infer the behavior of the tail of the distribution.

Finally, note that unless the chosen distribution has a finite bound (either a maximum for extremes in the upper tail or a minimum for the lower tail), one will not obtain $\widehat{p} = 0$.

3.3. Extreme value analysis (EVA)

A compromise between the fully parametric and nonparametric approaches is to use extreme value techniques developed in the statistical literature; Coles (2001) provides an excellent overview. This approach applies when the event of interest is sufficiently far in the tail of the distribution of R. The approach involves fitting a three-parameter statistical distribution only to the extreme observations. While the approach takes a parametric form, statistical theory provides strong theoretical support for the particular distributional form. We provide an overview of these methods, including a discussion of the statistical and implementational challenges of using the methods for annual or seasonal extremes in Supplemental Material C.

4. Estimating uncertainty in the risk ratio

In this section we consider several methods for estimating uncertainty using frequentist confidence intervals. Our focus is on confidence intervals for RR, and we introduce each method in the simplest context of the nonparametric estimator (Section 3.1) before describing how the method could be used in the context of EVA. After describing the methods we carry out a simulation study to assess which performs best.

4.1. The frequentist interpretation of uncertainty about p and RR

As discussed briefly in Section 2, we *cannot* generate a distribution of p or of $\text{RR} = p_F/p_C$ using the methods discussed here, as this is not part of the frequentist statistical framework used here. $\widehat{\text{RR}}$ has a distribution, called the sampling distribution, that is induced by the distributions of $\widehat{p}_F$ and $\widehat{p}_C$, but it is not particularly useful to plot it, as it represents the variability of $\widehat{\text{RR}}$ around RR. The danger in presenting such a plot is that it will often be viewed incorrectly as representing the distribution of RR, interpreted in a Bayesian fashion even though it has not been derived in a Bayesian fashion. For example, suppose the sampling distribution has a long right tail. This indicates that the estimator, $\widehat{\text{RR}}$, might be much larger than the true value, RR, and suggests that the true value may be much smaller than the our estimate. But if this sampling distribution were incorrectly interpreted as a Bayesian posterior, one would conclude that the true value may be much larger than the point estimate. (However, in the case of a simple parametric model and large sample size, the Bayesian central limit theorem states that the Bayesian posterior distribution will be approximately the same as the frequentist sampling distribution (Gelman et al., 2013)).

Instead, the standard statistical approach to quantifying uncertainty is to estimate a confidence interval. For the sake of illustration we will discuss a 90% confidence interval here, but the ideas are the same for other levels of confidence. The basic principle of a 90% confidence interval is that it is a random interval that should include the true value of the RR in 90% of the datasets that we might observe/collect. It is a random interval because it varies depending on the data collected, while the true RR is considered to be fixed. Therefore given a procedure for producing a confidence interval we can assess the procedure based on its coverage probability (the probability it contains the true value), assessed using synthetic datasets developed to mimic the real data-generating mechanism. Our goal is to use a procedure for calculating a confidence interval that produces intervals as short as possible while still having the specified coverage probability of 90%.

4.2. Methods for calculating confidence intervals

In this section, we explore several ways to estimate a confidence interval for the RR. We note that the statistical literature on estimating the RR (and the related odds ratio) in the biomedical literature is large and in particular considers a number of methods applied to binomial counts (see Fagerland et al. (2015) for an overview), and we consider some of those methods in addition to the bootstrap and a basic normal-theory method. Many of these methods are implemented in the climextRemes package available for R and Python, which builds upon the extRemes R package (Gilleland and Katz, 2011).

In general for probabilities (especially those near 0 or 1) and for ratios the sampling distribution of a given estimator is not a normal distribution, often being skewed. In this setting, working on the log scale often gives better statistical performance, with confidence intervals that come closer to having the desired coverage (e.g., Katz et al., 1978). One would compute the confidence interval for the quantity on the (natural) log scale (e.g., the log risk ratio) and then exponentiate the endpoints of the interval to get a confidence interval on the original scale.

4.2.1. Normal-theory confidence intervals

Basic statistical theory provides a standard confidence interval based on the standard error of the estimator when the sampling distribution of the estimator is approximately normally distributed. The approach is appropriate for larger sample sizes but can perform poorly for smaller sample sizes and when either p_F or p_C is close to zero. Furthermore, it involves some mathematical derivation, with use of the delta method in the context of the RR. For completeness we provide details in Supplemental Material D.1.

We turn next to the bootstrap, which aims to overcome some of these difficulties by relying on computation.

4.2.2. Bootstrap

The bootstrap is a widely-used, asymptotically-justified statistical tool for estimating the uncertainty in statistical estimates, particularly in cases where one cannot derive the standard error of or confidence interval for an estimator in closed form (Efron and Tibshirani, 1994; Davison and Hinkley, 1997). To introduce the bootstrap we present it to develop a confidence interval for p for simplicity. However the same techniques work on the log scale for $\log p$ and for logRR.

A common version of the bootstrap involves resampling values of R with replacement from the sample, $r_1, \ldots, r_{n_w}$ to generate n_b different bootstrap datasets, each of size n_w. With each bootstrap dataset we estimate p, giving us $\widehat{p}^{(1)}, \ldots, \widehat{p}^{(n_b)}$. Note that the method does *not* provide us with a distribution to represent uncertainty in p and should not be interpreted as a Bayesian posterior distribution for p (though there are connections between Bayesian methods and the bootstrap). Rather the sample $\widehat{p}^{(1)}, \ldots, \widehat{p}^{(n_b)}$ provides us with an estimate of the sampling distribution of $\widehat{p}$. The idea of the bootstrap is that the empirical distribution of the $\widehat{p}^{(i)}$ values around $\widehat{p}$ generally behaves similarly to the distribution of $\widehat{p}$ around p, which is the true sampling distribution, $f(\widehat{p}; p)$. If we knew the true sampling distribution, it would be simple to calculate a confidence interval. E.g., with normal data, we know that $\overline{x} \sim \text{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Given knowledge of this sampling distribution, we can analytically derive a 90% confidence interval for μ as $\overline{x}$ plus/minus 1.64 times the standard error, $\sigma/\sqrt{n}$, without needing to use the bootstrap. In the absence of a known sampling distribution, there are a variety of ways to use the $\widehat{p}^{(i)}$ values to estimate a bootstrap confidence interval for p:

- **Bootstrap interval using bootstrap standard error estimate**: use the empirical standard deviation of the $\widehat{p}^{(i)}$ values around their mean, $\overline{\widehat{p}}$. If we are happy to assume normality, we can use the empirical standard deviation of the $\widehat{p}^{(i)}$ values as the standard error estimate, which we denote $\widehat{se}$, to form a 90% confidence interval in the usual way,

$$(\hat{p} - 1.64 \cdot \hat{se}, \hat{p} + 1.64 \cdot \hat{se}). \quad (4)$$

- **Bootstrap percentile interval**: use percentiles of the empirical distribution of the $\hat{p}^{(i)}$ values to avoid the assumption of normality needed when using the bootstrap-based standard error estimate. This gives the following 90% interval, involving the 5th ($\hat{p}^{5\%}$) and 95th ($\hat{p}^{95\%}$) percentiles of the $\hat{p}^{(i)}$ values,

$$(\hat{p}^{5\%}, \hat{p}^{95\%}). \quad (5)$$

- **Basic bootstrap interval**: the interval also involves $\hat{p}^{5\%}$ and $\hat{p}^{95\%}$ and is

$$(\hat{p} - (\hat{p}^{95\%} - \hat{p}),\ \hat{p} - (\hat{p}^{5\%} - \hat{p})). \quad (6)$$

It is useful to consider the intuition behind why the upper quantile is involved in calculating the lower endpoint of the interval and vice versa, in contrast to the percentile method. The basic intuition is that if $\hat{p}^{95\%}$ is much larger than $\hat{p}$, this indicates the plausibility of estimated values that are much larger than the true value. Given this, our lower limit for p should be much lower than $\hat{p}$, because our actual estimate, $\hat{p}$, calculated using our single dataset may be much larger than the true value, p. Note that if the sampling distribution is symmetric, with $\hat{p} - \hat{p}^{5\%} = \hat{p}^{95\%} - \hat{p}$ then the percentile and basic intervals will be the same.

- **Studentized bootstrap interval**: this approach improves upon the basic interval by standardizing by an estimate of the standard error in each bootstrap sample. Let $z^{(i)} = (\hat{p}^{(i)} - \hat{p})/\hat{se}^{(i)}$ where $\hat{se}^{(i)}$ is a (possibly rough) estimate of the standard error of $\hat{p}^{(i)}$ that is calculated from the ith bootstrap sample. Let $z^{5\%}$ and $z^{95\%}$ be the 5th and 95th percentiles of the $z^{(i)}$ values. The studentized confidence interval is then

$$(\hat{p} - \hat{se} \cdot z^{95\%},\ \hat{p} - \hat{se} \cdot z^{5\%})$$

where $\hat{se}$ is the (possibly rough) standard error estimate based on the actual dataset.

- **Adjusted percentile (BCa) bootstrap interval**: this approach seeks to improve upon the percentile interval by estimating a transformation that brings the sampling distribution closer to normality (Davison and Hinkley, 1997).

Confidence intervals for RR In general, the ensemble members under the two scenarios are unrelated. If they were related, we would want to resample from the data in a way that reflected the relationship between the simulations in the two scenarios. Given the lack of relationship, a bootstrap procedure is to obtain n_b resampled datasets from each scenario and to randomly pair the datasets to get n_b pairs, from which one can calculate $\log \widehat{\mathrm{RR}}^{(1)}, \ldots, \log \widehat{\mathrm{RR}}^{(n_b)}$. This pertains to estimates obtained from any of the nonparametric, parametric, or EVA methods.

Drawbacks to the bootstrap An important difficulty with the bootstrap occurs when $\widehat{\mathrm{RR}} = \infty$ (the following discussion holds equivalently for $\widehat{\mathrm{RR}} = 0$). In this case in which $\hat{p}_C = 0$ the bootstrap fails because there is no variability in the data; all bootstrap datasets will have $\hat{p}_C^{(i)} = 0$. Another case is when some of the bootstrap samples have $\hat{p}_C^{(i)} = 0$ or $\hat{p}_F^{(i)} = 0$ and therefore $\log \widehat{\mathrm{RR}}^{(i)} \in \{-\infty, \infty\}$. In this case, the bootstrap standard error estimate (and therefore the bootstrap normal interval) cannot be calculated, while the ad hoc approach of removing such bootstrap samples before calculating confidence intervals using the various bootstrap methods has no clear justification and could affect performance of the resulting confidence interval. Finally when using EVA, it is not uncommon to be unable to estimate $\hat{p}_F$ or $\hat{p}_C$ because the optimization for the EVA parameters does not converge, and this often occurs with a subset of the bootstrapped datasets as well. It is unclear how to handle this, although if the number of times this occurs is small, ad hoc calculation of confidence intervals based on ignoring these values may not cause serious bias in the uncertainty quantification.

More generally than just in this context of estimating the RR, the percentile bootstrap method is known to perform poorly in practice (Hesterberg, 2015). However, while the logic behind the swapping of the percentiles in the basic bootstrap method relative to the percentile method suggests we might favor the basic bootstrap, both theoretical results (Davison and Hinkley, 1997; Hesterberg, 2015) and our simulation results (Section 5) show that the basic bootstrap also does not perform well. Furthermore, our simulation results suggest that all of the bootstrap-based methods have serious drawbacks.

This is not surprising, particularly for values of $\hat{p}_F$ or $\hat{p}_C$ near zero. The bootstrap is known to perform poorly when the resampling gives a discrete distribution with few values. An extreme manifestation of this occurs when $\hat{p}_N$ or $\hat{p}_A$ are zero. However, even when this does not occur, the sampling distribution often has probability mass at ∞, 0 and 0/0, and the bootstrap estimate of the sampling distribution of $\log \widehat{\mathrm{RR}} - \log \mathrm{RR}$ can provide a poor approximation to the true sampling distribution.

4.2.3. *Inverting a hypothesis test using the likelihood ratio statistic*

This approach is appealing because it can be used when $\hat{p}_F = 0$ (or $\hat{p}_C = 0$), providing a one-sided confidence interval that gives a lower (upper) bound on plausible values of RR, as done in Jeon et al. (2016). The usefulness of providing a bound on the RR should be apparent, because the bound allows us to assess the magnitude of the anthropogenic influence in light of uncertainty. The basic idea is to find a confidence interval by inverting a hypothesis test for the estimate of interest. A standard hypothesis test that is commonly applicable is a likelihood ratio test (Casella and Berger, 2002), which compares the likelihood of the data based on the maximum likelihood estimate to the likelihood of the data when constraining the parameters to represent a simpler setting in which a null hypothesis is assumed true. Confidence intervals based on likelihood ratio tests are widely used and well described in the statistical literature but have not been used in event attribution, so we provide details in Supplemental Material D.2.

4.2.4. *Binomial-based intervals*

For the case where the two probabilities in the risk ratio are estimated using the nonparametric approach, but not the EVA approach, the statistical literature provides a wide variety of methods to calculate a confidence interval for the risk ratio from independent binomial proportions, motivated by the vast number of analyses of risk ratios in the biomedical literature (Fagerland et al., 2015). Once again, a standard framework involves inverting a hypothesis test. Of the wide array of possibilities, some methods our literature search suggested to be promising are:

- Wilson's method: Hansen et al. (2014) propose to find an interval for the risk ratio by conditioning on the sum of events in the two scenarios to produce a binomially-distributed quantity and using an approximate confidence interval for a binomial probability proposed by (Wilson, 1927).
- Koopman's asymptotic score method: (Koopman, 1984) proposes to invert Pearson's chi square test. (Fagerland et al., 2015) found this method to perform reasonably well in simulations.
- Wang and Shan's method: Wang and Shan (2015) seek to improve upon existing methods by an inductive process to determine the ordering of how extreme different data values are with respect to providing evidence against the null hypothesis and then computing p-values by computing probabilities of data as or more extreme than observed, using the worst case value of the unknown nuisance

parameter. One can then invert the test to obtain a confidence interval. More details are in Supplemental Material D.3.

5. Simulation study to evaluate the methods

Given the variety of methods available, the potential for small sample sizes, the fact that p_F or p_C are often near zero, and the difficulty that the normality-based method and the bootstrap methods have when $\widehat{\text{RR}} = \infty$, we explored the performance of the methods using a simulation study. For simplicity and given the limitations of EVA discussed earlier, the simulation study focuses on the context of estimating the RR based on the nonparametric approach rather than EVA. The key factors that affect the statistical performance of the methods are the ensemble size, the true value of the RR, and the true probability of the event.

5.1. Design

For a given scenario, defined by the ensemble size (n), the RR, and the event probability (p_F), we generated 5000 synthetic datasets of y_F and y_C values, each from a binomial distribution with n ensemble members and probability of event p_F and $p_C = p_F/\text{RR}$, respectively. We considered $n \in \{25, 50, 100, 400\}$, where 50 is the minimum suggested in the C20C + Detection and Attribution project protocol (http://portal.nersc.gov/c20c) and 400 the number available for our case study. Our results focus on ensembles of size 100 as intermediate between the high uncertainty with smaller sizes and the fact that sizes as large as 400 will often not be available. We considered $\text{RR} \in \{1, 2, 4, 8, 16\}$ representing the range from a world where there is no anthropogenic influence to one with very strong influence. We considered $p_F \in \{0.01, 0.025, 0.05, 0.10, 0.20\}$ to represent a range in how extreme the event is. Given all possible combinations of values of n, RR, and p_F, we have 100 scenarios.

Note that while we only consider risk ratios greater than or equal to one, the results are equivalent for $p_C \in \{0.01, 0.025, 0.05, 0.10, 0.20\}$ with $\text{RR} \in \{1, 1/2, 1/4, 1/8, 1/16\}$ because the same calculations can be done for $\text{RR}^{-1} \equiv \frac{p_C}{p_F}$, followed by taking one over the resulting interval endpoints to get a confidence interval for RR.

We then applied the following methods to each synthetic dataset, thereby producing a set of 5000 confidence intervals for each method for each scenario. The methods were:

- Wilson's method (Section 4.2.4),
- Koopman's asymptotic score test inversion (Section 4.2.4),
- Wang-Shan exact test inversion (Section 4.2.4),
- normal-theory with delta method (Section 4.2.1),
- likelihood-ratio (LR) test inversion (Section 4.2.3),
- bootstrap normal (Section 4.2.2),
- percentile bootstrap (Section 4.2.2),
- basic bootstrap (Section 4.2.2),
- bootstrap-t (Section 4.2.2), and
- BCa bootstrap (Section 4.2.2).

We report results for 90% confidence intervals, considering the lower and upper endpoints separately because in real event attribution analyses, the focus will often be on the lower bound for the RR so as to assess the plausibility of no anthropogenic influence. Thus we focus on 95% one-sided confidence intervals. However we note that some of the methods were not developed as two separate one-sided intervals. Therefore, a two-sided interval may cover the true RR with the correct probability but when interpreted as two separate one-sided intervals not have the correct one-sided coverage probability.

We assessed performance using the following metrics:

- coverage probability of the intervals (proportion of times the 5000 intervals included the true RR),
- length of the interval (focusing on the magnitude of the lower bound), and
- proportion of intervals that could not be calculated because $\widehat{p}_F = 0$ or $\widehat{p}_C = 0$.

All code is available at https://bitbucket.org/lbl-cascade/event_attribution_uq_paper.

6. Results

Fig. 1 shows the coverage probability for a one-sided lower confidence interval, (RR_L, ∞). We see that for the LR, basic bootstrap, and bootstrap-t intervals, the intervals fail to include the true value at least 95% of the time. In contrast, the other methods are overly conservative (the intervals include the true value more than 95% of the time) for most or all values of RR. While values higher than 95% may sound appealing, they provide conservative results with intervals that are overly long and thereby increased uncertainty in estimating the RR.

Note that in this and other figures we omit results for the bootstrap normal interval. In almost all of the scenarios the standard deviation of the RR estimates in the bootstrap samples could not be calculated because one or more estimates were zero, infinity or 0/0, and omitting those estimates has no statistical justification.

Fig. 2 shows the coverage probability for a one-sided upper confidence interval, $(0, \text{RR}_U)$. Here the Wang-Shan method is conservative, the LR and Koopman methods show some undercoverage, and the bootstrap, Wilson, and delta methods have widely variable results, with substantial undercoverage in some cases.

Figure 3 shows the median value of RR_L, with higher values corresponding to shorter intervals and more statistical certainty about the value of RR. As expected, in general intervals are wider for smaller values of p_F (and therefore also of p_C), corresponding to fewer observed events. While the LR, basic bootstrap, and bootstrap-t methods have higher lower bounds, these are the methods showing undercoverage (particularly the bootstrap methods), so the shorter intervals are only obtained by violating the condition that coverage probability be correct. In contrast, the Koopman, Wilson, and Wang-Shan methods perform well, with the Koopman and Wilson approaches showing shorter intervals for smaller values of p_F.

Finally, Fig. 4 shows the proportion of simulated datasets for which a lower confidence bound could not be calculated, illustrating how commonly the bootstrap and delta methods fail, particularly for larger values of RR and lower values of p_F, which correspond to scenarios with fewer events.

Figs. 7–10 in the Supplemental Material A show coverage results for $n = 400$. In general these results are similar to those for $n = 100$, but for $n = 400$ we see that coverage probability is generally closer to 95%, as we would expect with increasing sample size. Results for $n \in \{25, 50\}$ were produced but are not shown; the results are similar quantitatively in terms of coverage and qualitatively in terms of the median bound values, but as expected show smaller lower bounds and higher proportions of datasets in which the bound(s) could not be calculated.

In summary, the bootstrap methods perform poorly. The LR method, which can be used both when analyzing binomial counts and when using EVA, provides some advantages, in particular its ability to provide intervals in most situations, including when $\widehat{p}_C = 0$. In contrast, the bootstrap intervals require $\widehat{p}_C > 0$ and do not provide meaningful intervals when too many of the bootstrap samples result in $\widehat{p}_C^{(i)} = 0$.

However, the LR interval shows undercoverage; in contrast, if one is using binomial counting, then the Koopman and Wang-Shan methods are possibilities and avoid the undercoverage of the LR method. However, the upper endpoint of Koopman shows some undercoverage and both methods can be too conservative, leading to overly-long intervals. Thus there is a tradeoff here, and a user may wish to consider how conservative they wish to be in their analysis. The Wilson method performs similarly

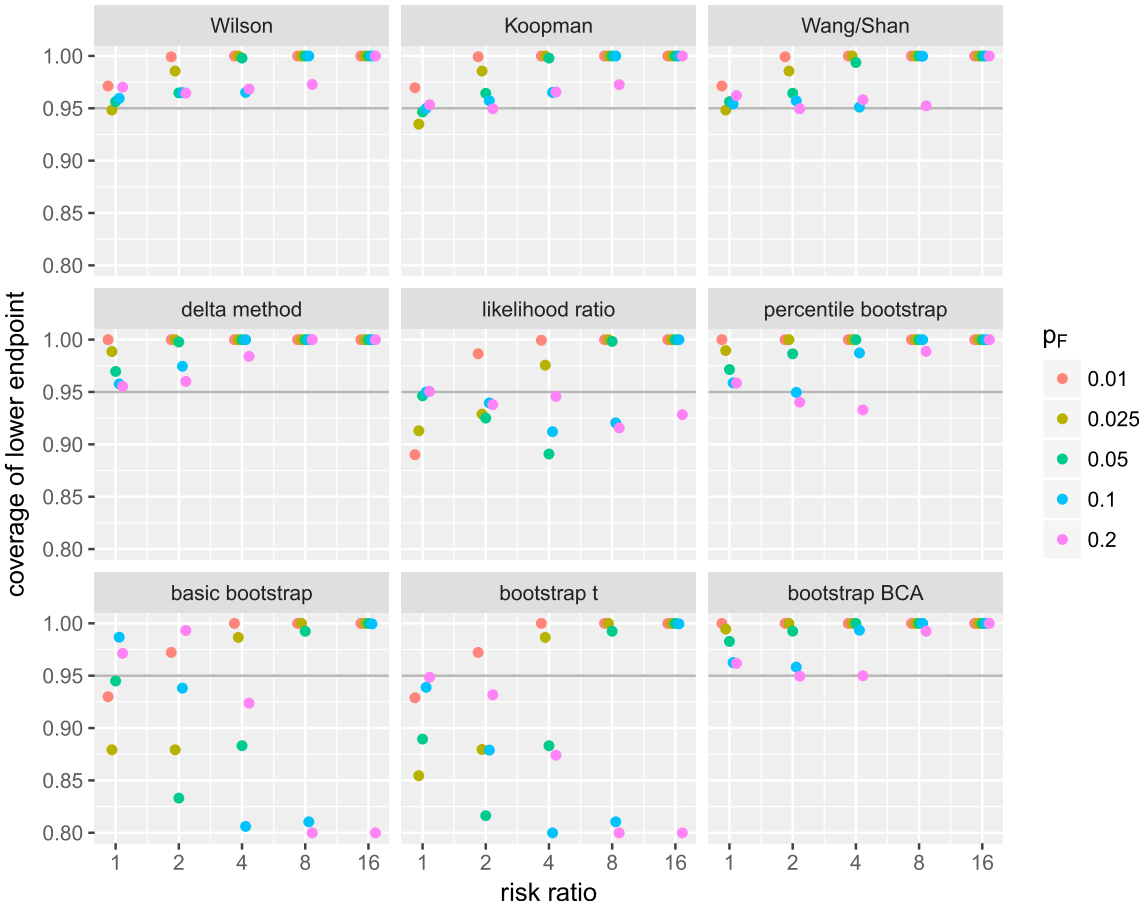

Fig. 1. Coverage probability of 95% lower confidence bound for $n = 100$ for various methods and values of RR and p_F. Values of 0.95 are optimal, while values less than 0.95 indicate undercoverage and values greater than 0.95 indicate conservativeness (overcoverage). Values lower than 0.8 are set to 0.8 for display purposes. Simulations in which the lower bound could not be computed were excluded.

to the Koopman method for the lower endpoint but has extreme undercoverage for the upper endpoint. The Koopman method is easy to compute, although the endpoints for Wang-Shan (whose calculation is computationally-intensive) can be pre-computed and saved as a lookup table.

7. Case study: Texas drought/heatwave

7.1. The event

Our case study focuses on the heat and dryness over the US state of Texas in the summer of 2011, the hottest and driest (in terms of precipitation deficit) on record dating back over a century (Rupp and Mote, 2012; Hoerling et al., 2013). Rupp and Mote (2012) performed a probabilistic event attribution analysis, but with counterfactual conditions estimated from previous years with similar anomalous temperature patterns in the Pacific Ocean (but cooler ocean temperatures generally). They concluded that anthropogenic emissions had increased the chance of the anomalously low rainfall and especially of the anomalous heat over Texas. While Hoerling et al. (2013) also found evidence for an anthropogenic contribution to the anomalous heat, they did not see evidence of a contribution to the precipitation deficit. Revealing possible inconsistencies in methods, however, these estimates based on climate models are at odds with the lack of an observed long-term summer warming over Texas (Stone et al., 2013; Hoerling et al., 2013). In this case study, we follow Rupp and Mote (2012) in considering March–August growing season temperature and rainfall averaged over the state of Texas, though the methods discussed in this work apply to longer or shorter periods.

7.2. Data and methods

We use data from CAM5.1 contributed to the C20C + Detection and Attribution Project (Folland et al., 2014). CAM5.1 simulates the processes of the atmosphere and land surface, given prescribed radiative and ocean surface conditions, run at a spatial resolution of approximately 1° (Neale et al., 2012). Simulations have been run under a factual scenario of observed boundary conditions (e.g., greenhouse gas concentrations, aerosol burdens, and sea surface temperatures) and under a counterfactual scenario with greenhouse gas and other radiative boundary conditions set to year-1855 values and ocean conditions set to the benchmark estimate of Stone and Pall. (2018). This removes a spatio-temporal pattern of anthropogenic warming but preserves month-to-month and year-to-year anomalous variability. We use data from 50 simulations of both scenarios covering a 1961–2010 reference period, 100 simulations (including the longer 50) covering the 1997–2010 period, and 400 simulations (including the longer 100) covering 2011–2013 (Angélil et al., 2017).

While multiple observational datasets are available, we use the CRU-TS-3.22 observationally-based data set (Harris et al., 2014). We work with anomalies of both observations and model output against the 1961–2010 period, by subtracting the historical mean for temperature and dividing by it for precipitation. For the model simulations, the factual scenario reference using the 50 simulations covering the full 1961–2010 period is used for estimating anomalies for both of the scenarios.

When using EVA we set the threshold, u, as the 90th percentile for temperature of the values being analyzed (both scenarios). For precipitation we consider the 20th percentile.

All code and data are available at https://bitbucket.org/lbl-cascade/

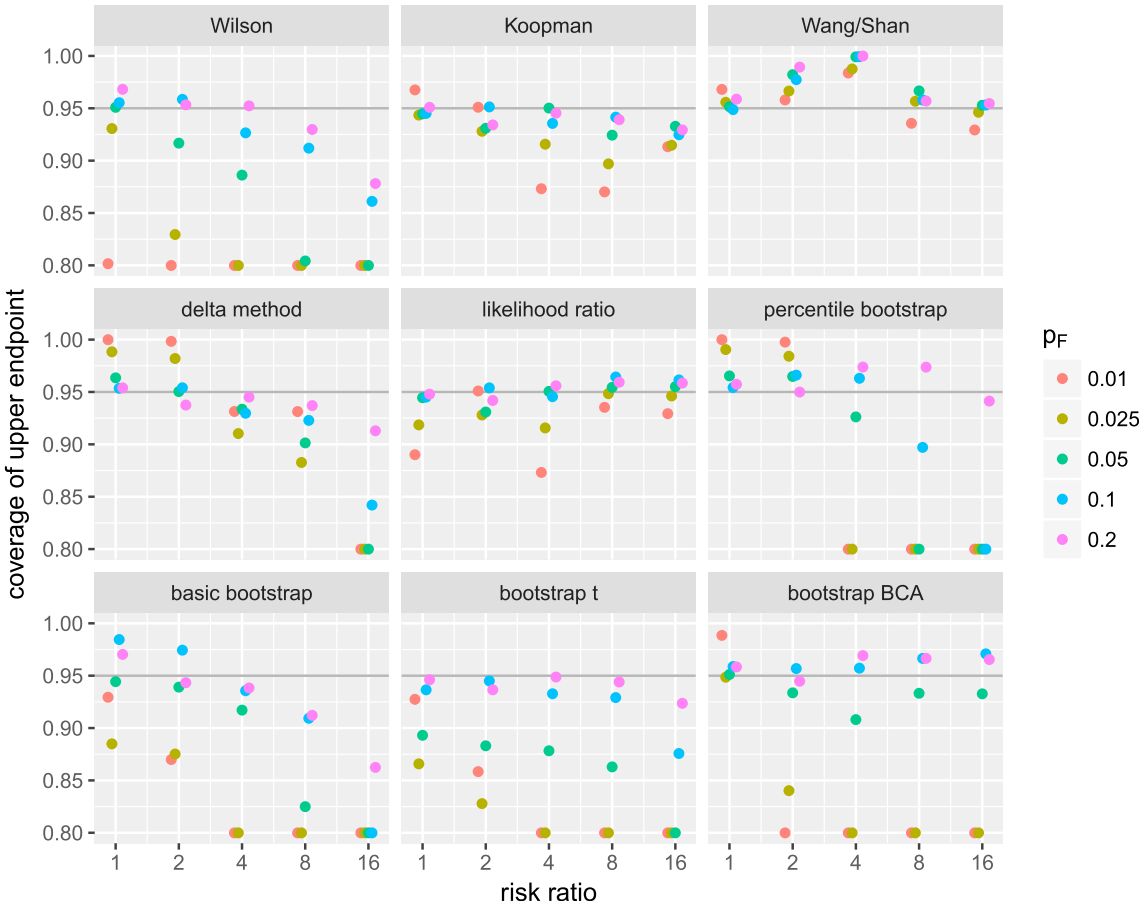

Fig. 2. Coverage probability of 95% upper confidence bound for $n = 100$ for various methods and values of RR and p_F. Values of 0.95 are optimal, while values less than 0.95 indicate undercoverage and values greater than 0.95 indicate conservativeness (overcoverage). Values lower than 0.8 are set to 0.8 for display purposes. Simulations in which the upper bound could not be computed were excluded.

event_attribution_uq_paper. We use the climextRemes R package version 0.2.0 (also available for Python) to estimate risk ratios and uncertainty, in some cases based on binomial counts and in others on EVA.

7.3. Uncertainty analysis for temperature

Fig. 5a shows the distribution of temperature anomalies for 2011 for the two scenarios.

The actual event of a 2.62 °C anomaly in 2011 is very extreme relative to both the factual and counterfactual distributions, particularly so with respect to the counterfactual. For the binomial approach, the estimated RR is ∞. Using the likelihood ratio-based confidence interval, we have a one-sided 95% confidence interval (CI) of (1.04, ∞), while using the Koopman method we have (0.74, ∞); both are quite uncertain because of how extreme the event is. Here EVA provides us with the ability to use the information in the sample more effectively because the event is so extreme. The RR estimate is still ∞, with $\widehat{p}_A = 0.0067$ and $\widehat{p}_N = 0$, but the 95% one-sided CI using the likelihood ratio approach is (12.8, ∞), providing strong evidence for a large RR.

Next, note that for less extreme event definitions, the event can be fairly common in the factual scenario and still not observed or very uncommon in the counterfactual. Given that the event is not extreme in the factual, EVA is not appropriate for that scenario, and we focus on the results based on the binomial approach. Table 1 shows results for a variety of definitions of the event. Note that we report a two-sided 90% CI by considering two one-sided 95% intervals.

Note that the factual distribution is shifted so substantially relative to the counterfactual distribution that the RR estimates are very large for extreme events and necessarily less so for less extreme events because $\widehat{p}_F$ has an upper bound at 1 and as the event becomes less extreme, $\widehat{p}_C$ increases and is no longer negligible.

Regardless, the evidence is strong that the probability of an extreme heatwave is much greater with anthropogenic influence than without, and for events defined by anomaly values larger than one, we conclude the risk ratio is at least 10, having accounted for sampling uncertainty. An important caveat is of course whether the model is able to capture the climatology of the event of interest, which is outside the scope of this work.

7.4. Uncertainty analysis for precipitation

Fig. 5b shows the distribution of precipitation (relative) anomalies for 2011 for the two scenarios. For the actual event, an anomaly value of 0.40 (i.e., 40% as much precipitation as the historical mean), we have zero exceedances in both scenarios. Furthermore EVA is of limited use as we get $\widehat{p}_F = 0.000088$ and $\widehat{p}_C = 0$, with high uncertainty: the likelihood-ratio based lower bound for RR is less than 0.01.

Given this, we consider several values for the event definition that are less extreme than the actual drought. Fig. 6 shows results from both the binomial count approach and EVA. In general we see evidence for a RR greater than one, with our best estimate of a RR of about two at a variety of event definitions. However the uncertainty in the lower bound at more extreme event definitions limits our ability to make a more robust attribution statement.

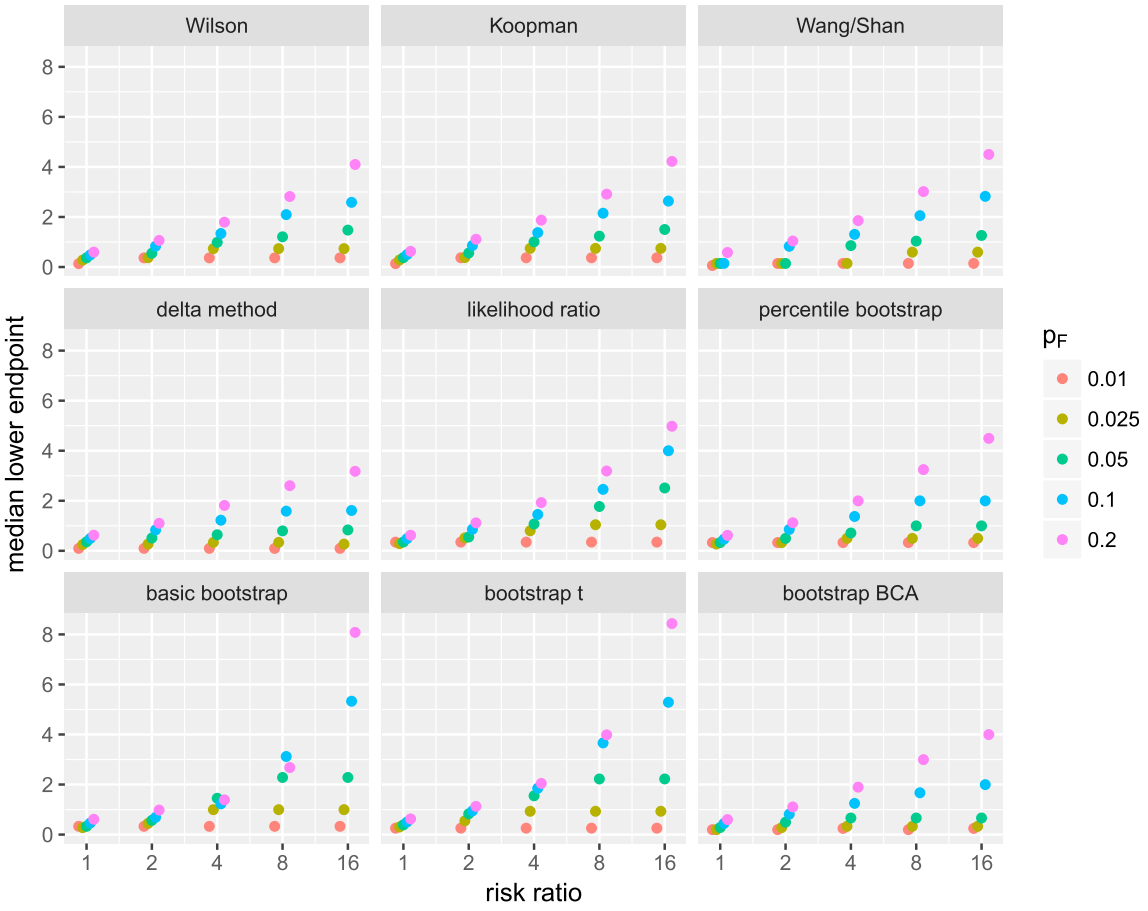

Fig. 3. Median value of lower confidence bound for $n = 100$ for various methods and values of RR and p_F. Higher values are better as they correspond to shorter confidence intervals. Simulations in which the lower bound could not be computed were excluded.

8. Discussion and recommendations

We have presented a statistical framework for estimating the risk ratio to quantify anthropogenic influence on extreme events, focusing on proposing, implementing, and evaluating standard frequentist statistical methods for accounting for sampling uncertainty in event attribution. Our methodological recommendations are as follows:

1 When using binomial counting, we recommend either the Koopman or Wang-Shan methods. If one is interested in the upper endpoint as well as the lower endpoint for small sample sizes, then Koopman does not preserve coverage probability, and we suggest Wang-Shan. These methods can be conservative, so the likelihood ratio approach might also be considered despite it producing intervals that can be too short.
2 When using EVA, only the normal-theory, likelihood ratio, and bootstrap methods are feasible. We recommend the likelihood ratio approach to calculate confidence intervals for the RR. This approach can provide a CI even when the estimate of the RR is infinity and avoids the problem that the bootstrap can fail to varying degrees when bootstrapped values of the RR estimate cannot be calculated or are infinity.
3 With the likelihood ratio approach, confidence intervals can be too short in a variety of circumstances and should be interpreted cautiously.
4 While appealing in its simplicity, bootstrap methods can perform poorly for quantifying uncertainty in RR, particularly when either p_F or p_C are near zero.
5 Bootstrapping provides a methodology to calculate confidence intervals, not Bayesian probability statements/intervals about RR. Unless an explicit Bayesian analysis is done, researchers should provide confidence intervals as their measure of uncertainty and avoid plotting the bootstrap distribution, as it represents the sampling distribution for $\widehat{RR}$ not the distribution of RR.
6 Although we have not investigated its performance in the simulation study, EVA may provide estimates that use the data more efficiently than binomial counting in some cases, particularly when sample sizes are not too small and the event definition is fairly extreme under both scenarios. In this situation, the small counts of extreme events lead to high uncertainty for the binomial counting approach, but EVA can borrow information from data values below the event definition cutoff. However, with small sample sizes, EVA suffers from having few observations with which to fit the distribution and often from difficulties in numerical optimization.
7 EVA is not appropriate if the event is not extreme in the scenario being analyzed (e.g., often the case for cold events in a natural counterfactual scenario). Furthermore simple binomial counts are effective and straightforward as sample sizes increase provided the event is not too rare in at least one of the scenarios. EVA and binomial counting could be used in the same analysis when the event is rare in one scenario and not the other. However, in this case it is clear that the RR is far from zero and there would be limited benefit from reducing uncertainty in the probability for the scenario in which the event is rare, so analysis using binomial counting for both scenarios should be very effective.

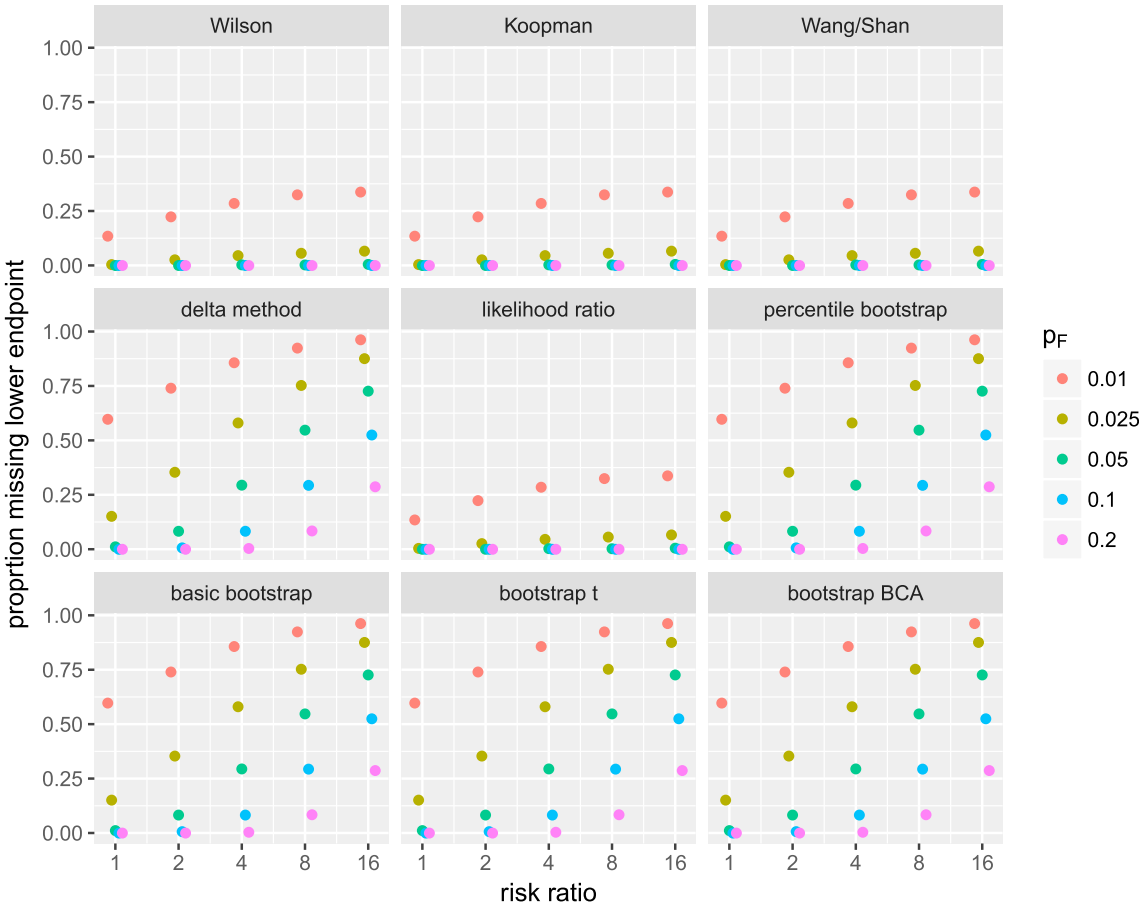

Fig. 4. Proportion of simulated datasets for which lower confidence bound could not be calculated for $n = 100$. For the Wilson, Koopman, Wang-Shan, and likelihood ratio methods this occurs only when no events occur in both scenarios.

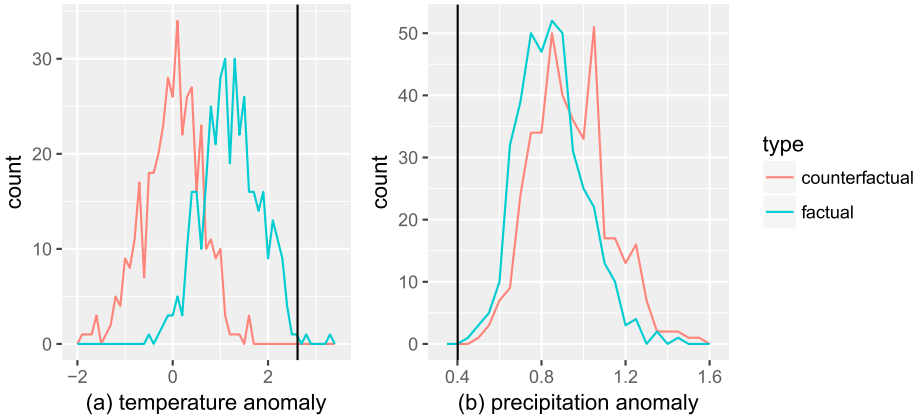

Fig. 5. Histograms of temperature (degrees Celsius) (a) and precipitation (b) anomalies for March–August 2011 over Texas from 400-member ensembles, with actual event indicated by black line.

Bayesian methods can also be useful and can be used to account for sampling uncertainty. Perhaps most usefully, a Bayesian perspective is probably the only option for accounting for uncertainty from uncertain boundary conditions (including uncertain counterfactual conditions in conditional attribution studies) and model parametric and structural uncertainty. For example, if one has an ensemble of model simulations based on varying the parameters of the climate model, this is most naturally seen as having drawn from a (prior) distribution over model parameters. This is naturally viewed in a Bayesian context but is difficult to conceptualize in a frequentist statistics framework as variation in the data that one might observe under the hypothetical of repeating an experiment. Similarly, the use of multiple climate models and the possibility of an ensemble of simulations of possible aerosol forcing or other forcings could be considered in a Bayesian framework (e.g., Smith et al., 2009). Critically, unlike with sampling variability, additional simulations do not reduce uncertainty from these sources or give results that converge in a frequentist statistical context to the true RR. In fact, from a statistical perspective one might characterize these uncertainties as bias

Table 1
RR estimates and 90% CI for a variety of event definitions based on binomial count approach. The last three event definitions are based on the quantiles of the CRU observations.

event definition (°C)	Number factual/ counterfactual exceedances	$\widehat{RR}$	Koopman CI	LRT CI
2.62 (actual event)	2/0	∞	(0.74, ∞)	(1.04, ∞)
2.0	43/0	∞	(16, ∞)	(31, ∞)
1.5	129/3	43	(17, 108)	(19, 133)
1.03 (1 in 20 year)	245/11	22	(14, 36)	(14, 38)
0.73 (1 in 10 year)	314/40	7.9	(6.1, 10.1)	(6.2, 10.2)
0.43 (1 in 5 year)	357/90	4.0	(3.4, 4.6)	(3.4, 4.7)

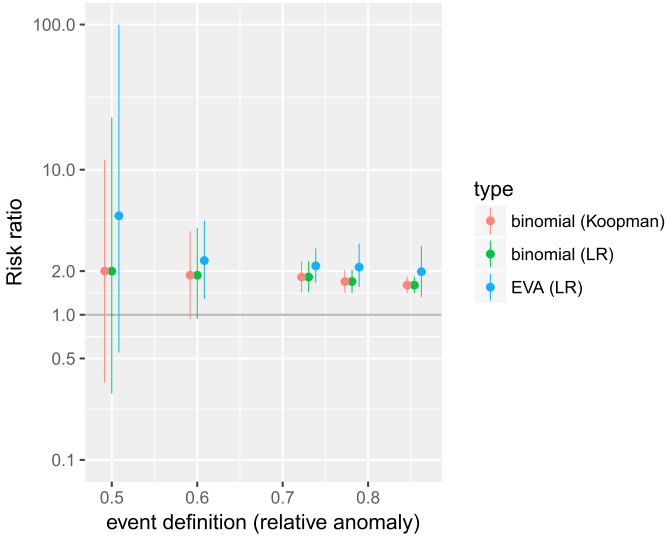

Fig. 6. Estimated risk ratio and 90% confidence intervals for both binomial counts and EVA for various definitions of the event in terms of March–August precipitation anomaly over Texas (so lower values are more extreme). The three event definitions to the right are the 1 in 20, 1 in 10 and 1 in 5 year events based on the CRU observations. The upper bound for EVA for the event of 0.5 is greater than 100 but truncated at 100 for plotting purposes. Note that for less extreme event definitions, EVA is less appropriate as EVA relies on the cutoff being in the tail of the distribution.

rather than variance. Given that conditioning on the ensemble output has limitations in terms of quantifying these uncertainties, any statistical treatment of these uncertainties is likely to represent some form of sensitivity analysis or subjective Bayesian analysis. Consideration of how to quantify parametric and structural uncertainty is an active area of climate research generally (e.g., Knutti et al., 2010).

One common question in event attribution analyses is how to define the event, which is often based on choosing the cutoff beyond which an 'extreme event' is considered to have occurred. When the motivation for a study is damage to society, one may be able to choose a specific cutoff (or small range of cutoffs) to use. If analysis is motivated by an event occurring and the cutoff is determined based on observations, one might consider the definition to be a source of uncertainty. However, our perspective (originally developed in Jeon et al. (2016)) is that this is most naturally considered as a sensitivity analysis, reporting the RR and uncertainty for a variety of cutoff values, as done in our case study and in Angélil et al. (2017) and Pall et al. (2017). In some cases, such as temperature in our Texas example, while the results vary somewhat with the event definition, the lower bound of the confidence interval provides evidence for a robust attribution statement in light of uncertainty.

Acknowledgments

We thank Oliver Angélil for contributing the C20C + D&A model simulations used in the analysis of the Texas event and Weizhen Wang for code to implement the method of Wang and Shan (2015). This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 as part of their Regional & Global Climate Modeling (RGCM) Program within the Calibrated And Systematic Characterization Attribution and Detection of Extremes Scientific Focus Area (CASCADE SFA).

Appendix A. Supplemental material

Supplemental material related to this article can be found at https://doi.org/10.1016/j.wace.2018.01.002.

References

Allen, M., 2003. Liability for climate change. Nature 421, 891–892.
Angélil, O., Perkins-Kirkpatrick, S., Alexander, L.V., Stone, D., Donat, M.G., Wehner, M., Shiogama, H., Ciavarella, A., Christidis, N., 2016. Comparing regional precipitation and temperature extremes in climate model and reanalysis products. Weather and Climate Extremes 13, 35–43.
Angélil, O., Stone, D., Wehner, M., Paciorek, C., Krishnan, H., Collins, W., 2017. An independent assessment of anthropogenic attribution statements for recent extreme temperature and rainfall events. J. Clim. 30, 5–16.
Casella, G., Berger, R., 2002. Statistical Inference. Duxbury Press, Pacific Grove, California.
Coles, S.G., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer Verlag, New York.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press.
Fagerland, M.W., Lydersen, S., Laake, P., 2015. Recommended confidence intervals for two independent binomial proportions. Stat. Meth. Med. Res. 24 (2), 224–254.
Folland, C., Stone, D., Frederiksen, C., Karoly, D., Kinter, J., 2014. The international CLIVAR climate of the 20th century plus (C20C+) project: report of the sixth workshop. CLIVAR Exchanges 19, 57–59.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. Chapman & Hall/CRC.
Gilleland, E., Katz, R.W., 2011. New software to analyze how extremes change over time. Eos, Transactions American Geophysical Union 92 (2), 13–14.
Hansen, G., Auffhammer, M., Solow, A.R., 2014. On the attribution of a single event to climate change. J. Clim. 27, 8297–8301.
Harris, I., Jones, P.D., Osborn, T.J., Lister, D.H., 2014. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 dataset. Int. J. Climatol. 34, 623–642.
Herring, S.C., Hoell, A., Hoerling, M.P., Kossin, J.P., C. J. Schreck III, Stott, P.A., 2016. Explaining extreme events of 2015 from a climate perspective. Bull. Am. Meteorol. Soc. 97 (12), S1–S145.
Herring, S.C., Hoerling, M.P., Kossin, J.P., Peterson, T.C., Stott, P.A., 2015. Explaining extreme events of 2014 from a climate perspective. Bull. Am. Meteorol. Soc. 96 (12), S1–S172.
Herring, S.C., Hoerling, M.P., Peterson, T.C., Stott, P.A., 2014. Explaining extreme events of 2013 from a climate perspective. Bull. Am. Meteorol. Soc. 95 (9), S1–S104.
Hesterberg, T.C., 2015. What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. Am. Statistician 69 (4), 371–386.
Hoerling, M., Kumar, A., Dole, R., Nielsen-Gammon, J.W., Eischeid, J., Perlwitz, J., Quan, X.-W., Zhang, T., Pegion, P., Chen, M., 2013. Anatomy of an extreme event. J. Clim. 26, 2811–2832.
Jeon, S., Paciorek, C.J., Wehner, M.F., 2016. Quantile-based bias correction and uncertainty quantification of extreme event attribution statements. Weather and Climate Extremes 12, 24–32.
Katz, D., Baptista, J., Azen, S., Pike, M., 1978. Obtaining confidence intervals for the risk ratio in cohort studies. Biometrics 34, 469–474.
Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G.A., 2010. Challenges in combining projections from multiple climate models. J. Clim. 23 (10), 2739–2758.
Koopman, P., 1984. Confidence intervals for the ratio of two binomial proportions. Biometrics 40, 513–517.
Lott, F., Stott, P., 2016. Evaluating simulated fraction of attributable risk using climate observations. J. Clim. 29, 4565–4575.
National Academies of Sciences, Engineering, and Medicine, 2016. Attribution of Extreme Weather Events in the Context of Climate Change. The National Academies Press.
Neale, R.B., Chen, C.-C., Gettelman, A., Lauritzen, P.H., Park, S., Williamson, D.L., Conley, A.J., Garcia, R., Kinnison, J.-F., Lamarque, D., Marsh, D., Mills, M., Smith, A.K., Tilmes, F., Vitt, S., Morrison, H., Cameron-Smith, P., Collins, W.D., Iacono, M.J., Easter, R.C., Ghan, S.J., Liu, X., Rasch, P.J., Taylor, M.A., 2012. Description of the NCAR Community Atmosphere Model (CAM 5.0). Technical Report. NCAR Technical Note NCAR/TN-486+STR.

Pall, P., Aina, T., Stone, D.A., Stott, P.A., Nozawa, T., Hilberts, A.G.J., Lohmann, D., Allen, M.R., 2011. Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. Nature 470 (7334), 382–385.

Pall, P., Patricola, C., Wehner, M., Stone, D., Paciorek, C., Collins, W., 2017. Diagnosing conditional anthropogenic contributions to heavy Colorado rainfall in september 2013. Weather and Climate Extremes 17 (1–6). https://doi.org/10.1016/j.wace.2017.03.004.

Peterson, T.C., Hoerling, M.P., Stott, P.A., Herring, S.C., 2013. Explaining extreme events of 2012 from a climate perspective. Bull. Am. Meteorol. Soc. 94 (9), S1–S74.

Peterson, T.C., Stott, P.A., Herring, S., 2012. Explaining extreme events of 2011 from a climate perspective. Bull. Am. Meteorol. Soc. 93, 1041–1067.

Rupp, D.E., Mote, P.W., 2012. Did human influence on climate make the 2011 Texas drought more probable? Bull. Am. Meteorol. Soc. 93, 1052–1054.

Smith, R.L., Tebaldi, C., Nychka, D., Mearns, L.O., 2009. Bayesian modeling of uncertainty in ensembles of climate models. J. Am. Stat. Assoc. 104 (485), 97–116.

Stone, D.A., Allen, M.R., 2005. The end-to-end attribution problem: from emissions to impacts. Climatic Change 71, 303–318.

Stone, D.A., Paciorek, C.J., Prabhat, P. Pall, Wehner, M.F., 2013. Inferring the anthropogenic contribution to local temperature extremes. Proc. Natl. Acad. Sci. Unit. States Am. 110, E1543.

Stone, D.A., Pall, P., 2018. A benchmark estimate of the effect of anthropogenic emissions on the ocean surface (in prep). in prep, . http://portal.nersc.gov/c20c/pub/StoneDA_PallP_2017.pdf.

Stott, P.A., Allen, M.R., Christidis, N., Dole, R., Hoerling, M., Huntingford, C., Pall, P., Perlwitz, J., Stone, D.A., 2013. Attribution of weather and climate-related extreme events. In: Asrar, G.R., Hurrell, J.W. (Eds.), Climate Science for Serving Society: Research, Modelling and Prediction Priorities. Springer, pp. 307–337.

Stott, P.A., Stone, D.A., Allen, M.R., 2004. Human contribution to the European heatwave of 2003. Nature 432 (7017), 610–614.

Wang, W., Shan, G., 2015. Exact confidence intervals for the relative risk and the odds ratio. Biometrics 71 (4), 985–995.

Wilson, E., 1927. Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. 22, 209–212.

Further reading

Farrington, C.P., Manning, G., 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat. Med. 9 (12), 1447–1454.

Risser, M.D., Stone, D.A., Paciorek, C.J., Wehner, M.F., Angélil, O., 2017. Quantifying the effect of interannual ocean variability on the attribution of extreme climate events to human influence. Climate Dynamics 49 (9–10), 3051–3073.